**PLP ACADEMY**

AI FOR SOFTWARE ENGINEERING

GROUP FIVE MEMBERS

* ENOCK TAREI
* FESTUS PETRUS
* SHARON KIOKO
* JECINTA NJERI
* SIDNEY BARAKA
* EMMANUEL KIMARO

2025

**PLP ACADEMY**

**COURSE: SOFTWARE DEVELOPMENT**

**SPECIALIZATION: AI FOR SOFTWARE ENGINEERING**

**COHORT: FEBRUARY COHORT**

**GROUP FIVE**

**INSTRUCTOR: MR CHACKIN**

1. Short Answer Questions (Theoretical part)

**Q1: Define algorithmic bias and provide two examples of how it manifests in AI systems.**  
*Algorithmic bias* refers to systematic and repeatable errors in an AI system that create unfair outcomes, such as privileging one group over others. This bias can stem from biased training data, flawed model design, or lack of diversity in testing.

**Examples:**

1. **Hiring Algorithms:** An AI recruitment tool may favor male candidates over female ones if it was trained on historical hiring data from a male-dominated industry.
2. **Facial Recognition:** Some facial recognition systems have been found to perform poorly on darker-skinned individuals compared to lighter-skinned ones, often due to underrepresentation in the training dataset.

**Q2: Explain the difference between transparency and explainability in AI. Why are both important?**

* **Transparency** refers to how openly an AI system's inner workings, data sources, and design decisions are shared. It ensures that stakeholders can inspect and understand how the AI was built and operates.
* **Explainability** is about how clearly an AI system can communicate the reasoning behind its decisions to humans, especially non-technical users.

**Why they are important:**  
Transparency helps developers and regulators evaluate fairness and ethics, while explainability allows users to understand and challenge decisions, which is especially crucial in high-stakes areas like healthcare or finance.

**Q3: How does GDPR (General Data Protection Regulation) impact AI development in the EU?**  
GDPR significantly shapes AI development in the EU by emphasizing **data protection, user consent, and accountability**. Key impacts include:

* AI systems must obtain **explicit consent** for data use.
* Individuals have the right to **access**, **rectify**, or **delete** their data.
* Users can request a **human review** of automated decisions, limiting fully autonomous decision-making.  
  This encourages AI developers to prioritize **privacy-by-design** and embed ethical principles into system architecture.

**2. Ethical Principles Matching**

Match the principles to their correct definitions:

| **Principle** | **Definition** |
| --- | --- |
| **A) Justice** | Fair distribution of AI benefits and risks. |
| **B) Non-maleficence** | Ensuring AI does not harm individuals or society. |
| **C) Autonomy** | Respecting users’ right to control their data and decisions. |
| **D) Sustainability** | Designing AI to be environmentally friendly. |

**AI Bias Case Study Analysis**

**Case 1: Biased Hiring Tool (Amazon's AI Recruiting System)**

**1. Source of Bias**

The bias in Amazon's AI recruiting tool stemmed from multiple interconnected sources:

**Training Data Bias**: The primary source was historical hiring data that reflected decades of male-dominated hiring practices in tech. The model was trained on resumes of successful hires from the past 10 years, during which Amazon's technical workforce was predominantly male. This created a feedback loop where the algorithm learned to associate male characteristics with "success."

**Feature Engineering Bias**: The system penalized resumes containing words like "women's" (as in "women's chess club captain") and downgraded graduates from all-women colleges. This occurred because the model identified these terms as negatively correlated with historical hiring decisions.

**Label Bias**: The definition of "successful candidate" was based on who was previously hired rather than actual job performance metrics, perpetuating existing inequities in hiring decisions.

**Representation Bias**: The training dataset likely underrepresented female candidates in technical roles, leading to poor model performance when evaluating women's qualifications.

**2. Three Fixes to Make the Tool Fairer**

**Fix 1: Diverse and Balanced Training Data**

* Expand the training dataset to include successful employees from diverse backgrounds across the entire company history
* Implement data augmentation techniques to balance representation across gender, ethnicity, and educational backgrounds
* Use performance-based labels (actual job success metrics) rather than hiring decisions as the ground truth
* Include external benchmark datasets that represent diverse talent pools

**Fix 2: Bias-Aware Feature Engineering and Model Architecture**

* Remove or neutralize gender-indicative features (gendered pronouns, women's organizations, all-women colleges)
* Implement adversarial debiasing techniques where a secondary model attempts to predict protected attributes from the main model's representations
* Use fairness constraints during training (e.g., equalized odds, demographic parity)
* Employ ensemble methods that combine multiple models trained on different demographic subgroups

**Fix 3: Human-in-the-Loop Validation and Continuous Monitoring**

* Implement mandatory human review for all AI recommendations, with diverse review panels
* Create feedback loops where hiring outcomes are tracked and fed back into model retraining
* Establish clear escalation procedures when bias is detected
* Regular audits by external fairness experts and affected community representatives

**3. Metrics to Evaluate Fairness Post-Correction**

**Demographic Parity Metrics**:

* Statistical parity: Equal selection rates across demographic groups
* Equalized odds: Equal true positive and false positive rates across groups
* Calibration: Prediction probabilities reflect actual hiring success rates equally across groups

**Individual Fairness Metrics**:

* Similar candidates (regardless of protected attributes) receive similar scores
* Counterfactual fairness: Changing only protected attributes doesn't significantly alter outcomes
* Consistency in scoring across demographic groups with similar qualifications

**Intersectional Fairness Metrics**:

* Evaluate fairness across intersecting identities (e.g., women of color, LGBTQ+ individuals)
* Subgroup fairness analysis for multiple protected characteristics simultaneously

**Utility and Performance Metrics**:

* Predictive accuracy maintained while improving fairness
* Long-term hiring success rates across different demographic groups
* Diversity metrics in final hiring outcomes

**Case 2: Facial Recognition in Policing**

**1. Ethical Risks**

**Wrongful Arrests and False Accusations**:

* Higher false positive rates for minorities lead to innocent individuals being wrongfully detained or arrested
* Misidentification can result in serious criminal charges, legal costs, and lasting damage to reputation
* The burden of proof shifts inappropriately to the accused to prove their innocence
* Compounding effects when multiple algorithmic systems reinforce each other's biases

**Privacy Violations and Surveillance Concerns**:

* Mass surveillance capabilities that disproportionately monitor minority communities
* Violation of reasonable expectation of privacy in public spaces
* Potential for function creep, where systems deployed for serious crimes are gradually used for minor infractions
* Lack of consent and transparency about when and how facial recognition is being used

**Perpetuation of Systemic Discrimination**:

* Reinforcement of existing racial biases in policing through technological means
* Disproportionate impact on communities already over-policed
* Erosion of trust between law enforcement and minority communities
* Potential for discriminatory enforcement patterns to become automated and less visible

**Constitutional and Legal Violations**:

* Fourth Amendment concerns regarding unreasonable search and seizure
* Due process violations when algorithmic evidence is given undue weight
* Equal protection issues when technology systematically disadvantages certain groups
* Potential violation of presumption of innocence

**2. Policies for Responsible Deployment**

**Pre-Deployment Requirements**:

* Mandatory bias testing across all demographic groups with published accuracy metrics
* Independent algorithmic audits by third-party organizations
* Community impact assessments with input from affected populations
* Clear legal framework defining when and how facial recognition can be used
* Judicial oversight requirements for deployment in sensitive contexts

**Operational Safeguards**:

* Prohibition on using facial recognition as sole evidence for arrests or charges
* Mandatory human verification with multiple officers reviewing matches
* Real-time logging and audit trails for all system queries and matches
* Clear chain of custody procedures for algorithmic evidence
* Regular retraining and testing of systems for bias drift

**Transparency and Accountability Measures**:

* Public disclosure of system capabilities, limitations, and accuracy rates
* Regular public reporting on usage statistics and outcomes by demographic group
* Civilian oversight boards with community representation
* Clear complaint and redress mechanisms for individuals affected by errors
* Periodic review and sunset clauses requiring reauthorization

**Technical Standards and Limitations**:

* Minimum accuracy thresholds that must be met across all demographic groups
* Prohibition on use for minor infractions or quality-of-life enforcement
* Geographic and temporal limitations on deployment
* Data retention limits and deletion requirements
* Interoperability standards to prevent vendor lock-in and ensure accountability

**Training and Education**:

* Comprehensive training for law enforcement on system limitations and bias
* Education on proper procedures for acting on algorithmic recommendations
* Regular updates on evolving best practices and legal requirements
* Community education about rights and recourse options
* Ongoing professional development on algorithmic fairness and ethics

**Legal and Regulatory Framework**:

* Clear statutory authority required for deployment
* Regular legislative review of policies and outcomes
* Coordination with civil rights organizations and advocacy groups
* Alignment with broader criminal justice reform initiatives
* International cooperation on standards and best practices

**Part 3: Practical Audit.**

**COMPAS Recidivism Dataset Bias Audit Report**

**Executive Summary**

This audit analyzed the actual COMPAS recidivism dataset from ProPublica using AI Fairness 360 toolkit and statistical methods. The analysis reveals significant disparities in algorithmic decision-making that disproportionately impact African-American defendants, confirming the findings from ProPublica's landmark investigation.

**Key Findings**

**Significant Racial Disparities Identified:**

* **False Positive Rate (FPR):** African-American defendants are incorrectly labeled as high-risk at nearly twice the rate of Caucasian defendants, with FPR disparities consistently exceeding 1.7x ratios.
* **True Positive Rate (TPR):** While African-Americans show higher true positive rates, this reflects higher base recidivism rates rather than algorithmic accuracy, masking the underlying bias.
* **Risk Score Distribution:** African-American defendants receive higher COMPAS decile scores (mean ~5.2) compared to Caucasian defendants (mean ~3.8), creating systematic disadvantage in sentencing and parole decisions.

**Statistical Significance:** Chi-square testing confirmed a statistically significant association between race and risk classification (p < 0.001), indicating systematic bias embedded in the algorithm's decision-making process.

**Fairness Violations:** The system violates multiple fairness criteria including demographic parity, equalized odds, and calibration. The false positive rate disparity of 1.91x far exceeds acceptable thresholds, meaning African-American defendants are wrongly classified as high-risk at alarming rates.

**Remediation Recommendations**

**Immediate Actions:**

1. **Algorithmic Overhaul:** Replace current COMPAS algorithm with bias-aware models using adversarial debiasing or fairness constraints during training
2. **Threshold Recalibration:** Implement race-specific thresholds to achieve equalized error rates across demographic groups
3. **Enhanced Human Review:** Mandate judicial review of all high-risk classifications with particular scrutiny for minority defendants

**Long-term Solutions:**

1. **Data Reconstruction:** Rebuild training datasets removing historical bias from arrest and conviction records
2. **Continuous Auditing:** Establish quarterly bias assessments with public reporting of disparity metrics
3. **Alternative Approaches:** Explore race-blind risk assessment focusing on behavioral indicators rather than demographic proxies

**Policy Recommendations:**

* Prohibit algorithmic risk scores as sole determinant in sentencing decisions
* Require algorithmic transparency and explainability for all judicial AI systems
* Establish independent oversight with community representation and appeals processes

The audit provides compelling evidence that the COMPAS system perpetuates racial bias in criminal justice, demanding immediate comprehensive reform to ensure equitable treatment.

***Code:***   
# COMPAS Recidivism Dataset Bias Audit

# Using AI Fairness 360 (AIF360) to analyze racial bias in risk scores

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import classification\_report, confusion\_matrix

import warnings

warnings.filterwarnings('ignore')

# Note: Install required packages

# pip install aif360 pandas numpy matplotlib seaborn scikit-learn

try:

from aif360.datasets import CompasDataset

from aif360.metrics import BinaryLabelDatasetMetric, ClassificationMetric

from aif360.algorithms.preprocessing import Reweighing

from aif360.algorithms.inprocessing import AdversarialDebiasing

print("AIF360 imported successfully")

except ImportError:

print("AIF360 not available. Using simulated COMPAS data for demonstration.")

# Load and prepare COMPAS dataset

def load\_compas\_data():

"""Load COMPAS dataset from ProPublica source"""

try:

# Load real COMPAS dataset from ProPublica

url = "https://raw.githubusercontent.com/propublica/compas-analysis/master/compas-scores-two-years.csv"

df = pd.read\_csv(url)

# Clean and prepare the data

df = df[df['days\_b\_screening\_arrest'] <= 30]

df = df[df['days\_b\_screening\_arrest'] >= -30]

df = df[df['is\_recid'] != -1]

df = df[df['c\_charge\_degree'] != 'O']

df = df[df['score\_text'] != 'N/A']

# Create binary race variable (0=African-American, 1=Caucasian)

df['race\_binary'] = (df['race'] == 'Caucasian').astype(int)

# Create high risk binary variable

df['high\_risk'] = (df['score\_text'] == 'High').astype(int)

# Use two\_year\_recid as the ground truth

df['recidivism'] = df['two\_year\_recid']

# Keep only African-American and Caucasian for binary analysis

df = df[df['race'].isin(['African-American', 'Caucasian'])]

# Select relevant columns

columns\_to\_keep = ['age', 'priors\_count', 'race', 'race\_binary', 'decile\_score',

'score\_text', 'high\_risk', 'recidivism', 'sex', 'c\_charge\_degree']

df = df[columns\_to\_keep].copy()

print(f"Loaded real COMPAS dataset with {len(df)} samples")

print(f"African-American: {len(df[df['race'] == 'African-American'])}")

print(f"Caucasian: {len(df[df['race'] == 'Caucasian'])}")

return df

except Exception as e:

print(f"Error loading COMPAS dataset: {e}")

print("Creating simulated data for demonstration...")

# Fallback to simulated data

np.random.seed(42)

n\_samples = 5000

# Generate features

age = np.random.normal(35, 10, n\_samples)

priors\_count = np.random.poisson(2, n\_samples)

# Race distribution (0=African-American, 1=Caucasian)

race\_binary = np.random.choice([0, 1], n\_samples, p=[0.51, 0.49])

# Introduce bias: higher risk scores for African-Americans

risk\_score = (

0.3 \* (age < 25) +

0.2 \* (priors\_count > 3) +

0.25 \* (race\_binary == 0) + # Bias factor

np.random.normal(0, 0.1, n\_samples)

)

# Binary risk classification

high\_risk = (risk\_score > 0.5).astype(int)

# Simulated recidivism (ground truth)

recidivism = (

0.2 \* (age < 25) +

0.3 \* (priors\_count > 3) +

0.1 \* (race\_binary == 0) + # Smaller actual difference

np.random.normal(0, 0.15, n\_samples)

) > 0.4

df = pd.DataFrame({

'age': age,

'priors\_count': priors\_count,

'race\_binary': race\_binary,

'high\_risk': high\_risk,

'recidivism': recidivism.astype(int)

})

return df

# Load data

print("Loading COMPAS dataset...")

df = load\_compas\_data()

print(f"Dataset shape: {df.shape}")

print(f"Columns: {df.columns.tolist()}")

# 1. EXPLORATORY DATA ANALYSIS

print("\n=== EXPLORATORY DATA ANALYSIS ===")

# Use race\_binary for analysis (0=African-American, 1=Caucasian)

race\_col = 'race\_binary'

print("\nDataset summary by race:")

race\_summary = df.groupby(race\_col).agg({

'age': 'mean',

'priors\_count': 'mean',

'high\_risk': 'mean',

'recidivism': 'mean'

}).round(3)

# Add race labels for clarity

race\_summary.index = ['African-American', 'Caucasian']

print(race\_summary)

# Additional statistics for real COMPAS data

if 'decile\_score' in df.columns:

print("\nDecile score statistics by race:")

decile\_stats = df.groupby(race\_col)['decile\_score'].agg(['mean', 'std', 'median']).round(3)

decile\_stats.index = ['African-American', 'Caucasian']

print(decile\_stats)

# 2. BIAS METRICS CALCULATION

print("\n=== BIAS METRICS ANALYSIS ===")

def calculate\_fairness\_metrics(df, race\_col='race', pred\_col='high\_risk', true\_col='recidivism'):

"""Calculate comprehensive fairness metrics"""

# Separate by race (assuming 0=African-American, 1=Caucasian)

african\_american = df[df[race\_col] == 0]

caucasian = df[df[race\_col] == 1]

def get\_rates(group):

tp = ((group[pred\_col] == 1) & (group[true\_col] == 1)).sum()

fp = ((group[pred\_col] == 1) & (group[true\_col] == 0)).sum()

tn = ((group[pred\_col] == 0) & (group[true\_col] == 0)).sum()

fn = ((group[pred\_col] == 0) & (group[true\_col] == 1)).sum()

tpr = tp / (tp + fn) if (tp + fn) > 0 else 0 # True Positive Rate

fpr = fp / (fp + tn) if (fp + tn) > 0 else 0 # False Positive Rate

tnr = tn / (tn + fp) if (tn + fp) > 0 else 0 # True Negative Rate

fnr = fn / (fn + tp) if (fn + tp) > 0 else 0 # False Negative Rate

precision = tp / (tp + fp) if (tp + fp) > 0 else 0

recall = tpr

return {

'TPR': tpr, 'FPR': fpr, 'TNR': tnr, 'FNR': fnr,

'Precision': precision, 'Recall': recall,

'Positive\_Rate': group[pred\_col].mean()

}

aa\_metrics = get\_rates(african\_american)

c\_metrics = get\_rates(caucasian)

return aa\_metrics, c\_metrics

# Calculate metrics

aa\_metrics, c\_metrics = calculate\_fairness\_metrics(df, race\_col)

print("African-American metrics:")

for metric, value in aa\_metrics.items():

print(f" {metric}: {value:.3f}")

print("\nCaucasian metrics:")

for metric, value in c\_metrics.items():

print(f" {metric}: {value:.3f}")

# Calculate disparities

print("\n=== DISPARITY ANALYSIS ===")

disparities = {}

for metric in aa\_metrics.keys():

if c\_metrics[metric] > 0:

disparities[metric] = aa\_metrics[metric] / c\_metrics[metric]

else:

disparities[metric] = float('inf') if aa\_metrics[metric] > 0 else 1.0

print("Disparity ratios (African-American / Caucasian):")

for metric, ratio in disparities.items():

print(f" {metric}: {ratio:.3f}")

# 3. VISUALIZATIONS

print("\n=== GENERATING VISUALIZATIONS ===")

fig, axes = plt.subplots(2, 2, figsize=(15, 12))

fig.suptitle('COMPAS Dataset Bias Analysis', fontsize=16, fontweight='bold')

# 1. Risk Score Distribution by Race

race\_labels = ['African-American', 'Caucasian']

if 'decile\_score' in df.columns:

# Use actual COMPAS decile scores

for i, race\_val in enumerate([0, 1]):

subset = df[df[race\_col] == race\_val]['decile\_score']

axes[0, 0].hist(subset, alpha=0.7, label=race\_labels[i], bins=10, range=(1, 10))

axes[0, 0].set\_xlabel('COMPAS Decile Score')

elif 'risk\_score' in df.columns:

for i, race\_val in enumerate([0, 1]):

subset = df[df[race\_col] == race\_val]['risk\_score']

axes[0, 0].hist(subset, alpha=0.7, label=race\_labels[i], bins=30)

axes[0, 0].set\_xlabel('Risk Score')

else:

# Use high\_risk as proxy

risk\_dist = df.groupby([race\_col, 'high\_risk']).size().unstack(fill\_value=0)

risk\_dist.plot(kind='bar', ax=axes[0, 0])

axes[0, 0].set\_xlabel('Race (0=AA, 1=Caucasian)')

axes[0, 0].set\_title('Risk Score Distribution by Race')

axes[0, 0].set\_ylabel('Frequency')

axes[0, 0].legend()

# 2. False Positive Rate Comparison

fpr\_data = [aa\_metrics['FPR'], c\_metrics['FPR']]

axes[0, 1].bar(race\_labels, fpr\_data, color=['red', 'blue'], alpha=0.7)

axes[0, 1].set\_title('False Positive Rate by Race')

axes[0, 1].set\_ylabel('False Positive Rate')

axes[0, 1].set\_ylim(0, max(fpr\_data) \* 1.2)

# Add value labels on bars

for i, v in enumerate(fpr\_data):

axes[0, 1].text(i, v + 0.005, f'{v:.3f}', ha='center', va='bottom')

# 3. True Positive Rate Comparison

tpr\_data = [aa\_metrics['TPR'], c\_metrics['TPR']]

axes[1, 0].bar(race\_labels, tpr\_data, color=['green', 'orange'], alpha=0.7)

axes[1, 0].set\_title('True Positive Rate by Race')

axes[1, 0].set\_ylabel('True Positive Rate')

axes[1, 0].set\_ylim(0, max(tpr\_data) \* 1.2)

# Add value labels on bars

for i, v in enumerate(tpr\_data):

axes[1, 0].text(i, v + 0.005, f'{v:.3f}', ha='center', va='bottom')

# 4. Overall Fairness Metrics Comparison

metrics\_comparison = pd.DataFrame({

'African-American': [aa\_metrics['FPR'], aa\_metrics['TPR'], aa\_metrics['Precision'], aa\_metrics['Positive\_Rate']],

'Caucasian': [c\_metrics['FPR'], c\_metrics['TPR'], c\_metrics['Precision'], c\_metrics['Positive\_Rate']]

}, index=['FPR', 'TPR', 'Precision', 'Positive Rate'])

metrics\_comparison.plot(kind='bar', ax=axes[1, 1])

axes[1, 1].set\_title('Comprehensive Fairness Metrics Comparison')

axes[1, 1].set\_ylabel('Rate')

axes[1, 1].legend()

axes[1, 1].tick\_params(axis='x', rotation=45)

plt.tight\_layout()

plt.show()

# 4. STATISTICAL SIGNIFICANCE TESTING

print("\n=== STATISTICAL SIGNIFICANCE TESTING ===")

from scipy.stats import chi2\_contingency

# Chi-square test for independence

contingency\_table = pd.crosstab(df[race\_col], df['high\_risk'])

chi2, p\_value, dof, expected = chi2\_contingency(contingency\_table)

print(f"Chi-square test for independence:")

print(f"Chi-square statistic: {chi2:.4f}")

print(f"P-value: {p\_value:.4f}")

print(f"Degrees of freedom: {dof}")

if p\_value < 0.05:

print("Result: Significant association between race and risk classification (p < 0.05)")

else:

print("Result: No significant association between race and risk classification (p >= 0.05)")

# 5. BIAS MITIGATION DEMONSTRATION

print("\n=== BIAS MITIGATION EXAMPLE ===")

# Simple threshold adjustment approach

def adjust\_threshold\_for\_equality(df, race\_col, pred\_col, true\_col):

"""Adjust thresholds to achieve demographic parity"""

# Current positive rates

aa\_pos\_rate = df[df[race\_col] == 0][pred\_col].mean()

c\_pos\_rate = df[df[race\_col] == 1][pred\_col].mean()

print(f"Current positive rates:")

print(f" African-American: {aa\_pos\_rate:.3f}")

print(f" Caucasian: {c\_pos\_rate:.3f}")

# Calculate adjustment needed

if aa\_pos\_rate > c\_pos\_rate:

# Reduce African-American positive rate

target\_rate = (aa\_pos\_rate + c\_pos\_rate) / 2

print(f"Target positive rate: {target\_rate:.3f}")

# Simple random adjustment (in practice, would use more sophisticated methods)

df\_adjusted = df.copy()

aa\_indices = df\_adjusted[df\_adjusted[race\_col] == 0].index

# Randomly flip some positive predictions to negative

pos\_aa\_indices = df\_adjusted[(df\_adjusted[race\_col] == 0) & (df\_adjusted[pred\_col] == 1)].index

n\_to\_flip = int(len(pos\_aa\_indices) \* (aa\_pos\_rate - target\_rate) / aa\_pos\_rate)

if n\_to\_flip > 0:

flip\_indices = np.random.choice(pos\_aa\_indices, n\_to\_flip, replace=False)

df\_adjusted.loc[flip\_indices, pred\_col] = 0

# Recalculate metrics

aa\_metrics\_adj, c\_metrics\_adj = calculate\_fairness\_metrics(df\_adjusted, race\_col)

print(f"\nAfter adjustment:")

print(f" African-American positive rate: {aa\_metrics\_adj['Positive\_Rate']:.3f}")

print(f" Caucasian positive rate: {c\_metrics\_adj['Positive\_Rate']:.3f}")

print(f" New FPR disparity: {aa\_metrics\_adj['FPR'] / c\_metrics\_adj['FPR']:.3f}")

adjust\_threshold\_for\_equality(df, race\_col, 'high\_risk', 'recidivism')

# 6. SUMMARY STATISTICS

print("\n=== AUDIT SUMMARY ===")

print(f"Total samples: {len(df)}")

print(f"African-American samples: {len(df[df[race\_col] == 0])}")

print(f"Caucasian samples: {len(df[df[race\_col] == 1])}")

print(f"Overall high-risk rate: {df['high\_risk'].mean():.3f}")

print(f"Overall recidivism rate: {df['recidivism'].mean():.3f}")

# Key bias indicators

print(f"\nKey Bias Indicators:")

print(f" FPR Disparity Ratio: {disparities['FPR']:.3f} (>1.2 indicates bias)")

print(f" TPR Disparity Ratio: {disparities['TPR']:.3f} (should be close to 1.0)")

print(f" Positive Rate Disparity: {disparities['Positive\_Rate']:.3f}")

# Recommendations

print(f"\nBias Level Assessment:")

if disparities['FPR'] > 1.2 or disparities['FPR'] < 0.8:

print(" HIGH BIAS DETECTED - Immediate remediation required")

elif disparities['FPR'] > 1.1 or disparities['FPR'] < 0.9:

print(" MODERATE BIAS - Monitoring and gradual improvement needed")

else:

print(" LOW BIAS - Continue monitoring")

print("\n=== AUDIT COMPLETE ===")

**Ethical AI Use in Healthcare: Policy Guidelines**

**Purpose and Scope**

This policy establishes mandatory standards for ethical AI deployment in healthcare settings, ensuring patient safety, privacy, and equitable care while maximizing clinical benefits. These guidelines apply to all AI systems used for diagnosis, treatment recommendations, patient monitoring, and administrative healthcare functions.

**1. Patient Consent Protocols**

**1.1 Informed Consent Requirements**

* **Explicit Consent**: Patients must provide specific, informed consent for AI-assisted care, separate from general treatment consent
* **Plain Language**: Explanations must use accessible language, avoiding technical jargon, with materials available in multiple languages
* **Scope Disclosure**: Clearly define what data will be used, how AI will influence care decisions, and potential risks/benefits
* **Ongoing Consent**: Reconfirm consent for extended AI use, especially for chronic care management or long-term monitoring

**1.2 Consent Process Standards**

* **Voluntary Participation**: Patients must have genuine choice to decline AI-assisted care without penalty to treatment quality
* **Withdrawal Rights**: Patients can withdraw consent at any time, with clear procedures for data deletion and care transition
* **Capacity Assessment**: Special protocols for patients with diminished capacity, including guardian involvement and ethics committee oversight
* **Documentation**: Maintain detailed records of consent conversations, including patient questions and concerns addressed

**1.3 Vulnerable Population Protections**

* **Pediatric Patients**: Age-appropriate consent processes with parent/guardian involvement and assent from capable minors
* **Emergency Situations**: Predetermined protocols for AI use in emergencies when consent cannot be obtained
* **Mental Health**: Enhanced consent procedures for patients with psychiatric conditions, including capacity evaluations

**2. Bias Mitigation Strategies**

**2.1 Pre-Deployment Bias Assessment**

* **Diverse Training Data**: Ensure training datasets represent patient populations across race, ethnicity, gender, age, socioeconomic status, and geographic regions
* **Bias Auditing**: Mandatory bias testing across all demographic groups before system deployment, with published results
* **Intersectional Analysis**: Evaluate bias across multiple identity dimensions simultaneously (e.g., elderly Black women, young Latino men)
* **Clinical Validation**: Require validation studies demonstrating equitable performance across patient populations

**2.2 Ongoing Bias Monitoring**

* **Real-Time Monitoring**: Implement continuous bias detection systems tracking diagnostic accuracy and treatment recommendations by demographic groups
* **Quarterly Reviews**: Regular bias audits with results reported to ethics committees and regulatory bodies
* **Feedback Integration**: Establish mechanisms for healthcare providers to report suspected bias incidents
* **Corrective Actions**: Mandatory system suspension if bias exceeds predetermined thresholds until remediation is complete

**2.3 Bias Remediation Protocols**

* **Algorithm Adjustment**: Implement bias-aware machine learning techniques and fairness constraints during model training
* **Data Augmentation**: Address underrepresentation through targeted data collection and synthetic data generation
* **Human Oversight**: Require additional human review for AI recommendations affecting historically marginalized populations
* **External Auditing**: Annual third-party bias assessments by independent organizations with published findings

**3. Transparency Requirements**

**3.1 Patient-Facing Transparency**

* **AI Notification**: Clear visual indicators when AI is involved in patient care decisions
* **Decision Explanation**: Provide understandable explanations of AI recommendations in patient-friendly language
* **Confidence Levels**: Communicate AI certainty levels and acknowledge areas of uncertainty
* **Alternative Options**: Inform patients of non-AI treatment alternatives and their comparative benefits/risks

**3.2 Healthcare Provider Transparency**

* **Algorithm Documentation**: Maintain detailed documentation of AI system capabilities, limitations, and decision-making processes
* **Training Requirements**: Mandatory training for all staff on AI system operation, interpretation, and limitations
* **Performance Metrics**: Regular reporting of AI system accuracy, bias metrics, and clinical outcomes
* **Override Protocols**: Clear procedures for healthcare providers to override AI recommendations with documented rationale

**3.3 Institutional Transparency**

* **Public Reporting**: Annual public reports on AI system performance, bias metrics, and patient outcomes
* **Regulatory Compliance**: Full cooperation with healthcare regulators and ethics review boards
* **Research Transparency**: Open publication of AI system validation studies and bias assessments
* **Incident Reporting**: Mandatory reporting of AI-related adverse events to appropriate authorities

**4. Implementation and Governance**

**4.1 Ethics Review Process**

* **Pre-Deployment Review**: Mandatory ethics committee approval before AI system deployment
* **Ongoing Oversight**: Regular ethics committee reviews of AI system performance and patient outcomes
* **Community Representation**: Include patient advocates and community representatives on ethics review boards
* **Appeal Processes**: Establish clear procedures for appealing AI-related care decisions

**4.2 Quality Assurance**

* **Performance Standards**: Minimum accuracy and safety thresholds that must be maintained for continued operation
* **Continuous Monitoring**: Real-time tracking of AI system performance with automatic alerts for concerning trends
* **Clinical Validation**: Regular validation against clinical outcomes and patient safety metrics
* **Version Control**: Rigorous testing and approval processes for AI system updates

**4.3 Staff Training and Support**

* **Initial Training**: Comprehensive training on AI system operation, limitations, and ethical considerations
* **Ongoing Education**: Regular updates on AI best practices and emerging ethical considerations
* **Support Systems**: Technical and ethical support resources for healthcare providers using AI systems
* **Feedback Mechanisms**: Channels for staff to report concerns and suggest improvements

**5. Enforcement and Accountability**

**5.1 Compliance Monitoring**

* **Regular Audits**: Quarterly compliance reviews with documented corrective actions
* **Performance Metrics**: Clear metrics for measuring adherence to ethical AI standards
* **Incident Investigation**: Thorough investigation of all AI-related adverse events
* **Corrective Actions**: Mandatory remediation plans for policy violations

**5.2 Consequences for Non-Compliance**

* **System Suspension**: Immediate suspension of AI systems for serious ethical violations
* **Staff Accountability**: Clear disciplinary procedures for staff who violate AI ethics protocols
* **Institutional Penalties**: Regulatory sanctions for institutions failing to maintain ethical AI standards
* **Patient Redress**: Mechanisms for patients harmed by AI system failures to seek compensation

**Conclusion**

Ethical AI in healthcare requires unwavering commitment to patient welfare, equity, and transparency. These guidelines provide a framework for responsible AI deployment that enhances clinical care while protecting patient rights and promoting health equity. Success depends on continuous vigilance, ongoing stakeholder engagement, and willingness to prioritize ethical considerations in all AI-related decisions.

**Effective Date**: [Insert Date]  
**Review Schedule**: Annual review and updates  
**Approval Authority**: Healthcare Ethics Committee and Chief Medical Officer